

LARGE MODEL AGENTS: A SURVEY ON TECHNICAL EVOLUTION, MULTIMODAL FUSION, AND VERTICAL DOMAIN APPLICATIONS

ShouTong Huang, Yu Ma*

School of Electronic and Electrical Engineering, Ningxia University, Yinchuan 750021, Ningxia, China.

**Corresponding Author: Yu Ma*

Abstract: LLM-driven agent technology has achieved breakthrough progress, marking AI's leap from perceptual to actionable intelligence. This paper reviews four evolution stages: symbolic, reactive, RL, and LLM-driven agents. It analyzes core architecture including ReAct, memory, tool calling, and multimodal perception. For education and healthcare, we design a GLM-4-based multimodal educational assistant and a ReAct-based MedGuard medication assistant, validating effectiveness through experiments. Finally, we analyze six challenges (hallucination, alignment, security, planning, cost, interpretability) and propose directions: world models, RL fine-tuning, verifiable execution, edge deployment, and ethical frameworks.

Keywords: LLM agent; Multimodal fusion; GLM-4; ReAct; Education; Healthcare; Tool calling

1 INTRODUCTION

AI's history is one of expanding agent capabilities. From symbolic expert systems to reactive robots to RL systems, agents have gradually strengthened. However, traditional paradigms face bottlenecks: difficult knowledge acquisition, poor generalization, low sample efficiency. Since 2020, LLMs (GPT, GLM) provide a solid brain foundation [1]. LLM-based agents integrate perception, memory, planning, tools, and action, achieving the leap from language understanding to intelligent action [2].

Multimodal models (GLM-4V, CogView, CogVideo) further enable visual perception and cross-modal generation. Yet challenges remain: hallucination, alignment deviation, tool abuse risks, goal forgetting, high costs, and black-box decisions. This paper systematically reviews evolution, architecture, applications, and future directions. Contributions: (1) complete evolution review; (2) analysis of ReAct, memory, tools, multimodal fusion; (3) education and healthcare cases with experiments; (4) six-dimension challenge framework.

2 TECHNICAL FOUNDATION AND EVOLUTION

2.1 Four Stages

(1) Symbolic Agent 1950s-1980s: logic rules, knowledge graphs, STRIPS, MYCIN. Strong interpretability but knowledge acquisition bottleneck. (2) Reactive Agent 1980s-2000s: Brooks Subsumption [3], perception-action mapping. Real-time but no high-level reasoning. (3) RL Agent 2000s-2020: DQN, AlphaGo [4]. Superhuman on specific tasks but low sample efficiency. (4) LLM-Driven Agent 2020-present: prompting, CoT, tool use, memory. Landmarks: 2022 AutoGPT, 2023 ChatGPT+Plugins, 2024 Astra, 2025 MCP, see Figure 1 [5].

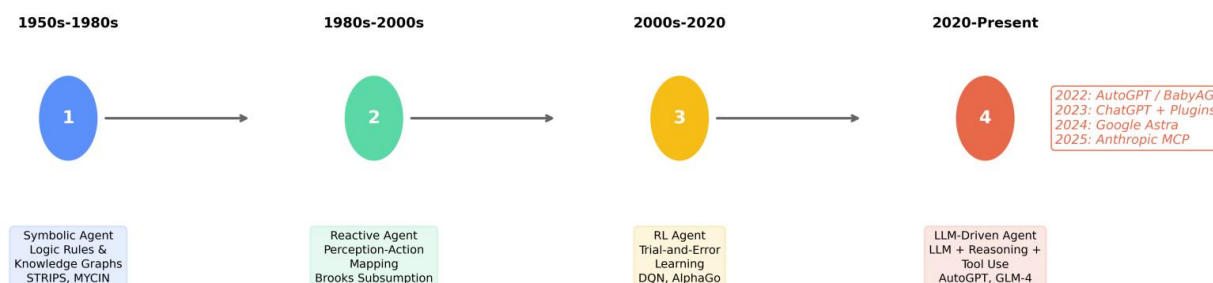


Figure 1 Evolution Roadmap of Large Model Agents

2.2 Multimodal Fusion

Early LLM agents relied on pure text. With CLIP, ViT, GLM-4V, agents gained image-text understanding. CogView, CogVideo enable generation [6]. Paths: encoder alignment, end-to-end multimodal LLMs, tool chain integration.

3 CORE ARCHITECTURE AND KEY TECHNOLOGIES

Mainstream systems adopt a perception-reasoning-memory-tools-security layered architecture, see Figure 2.

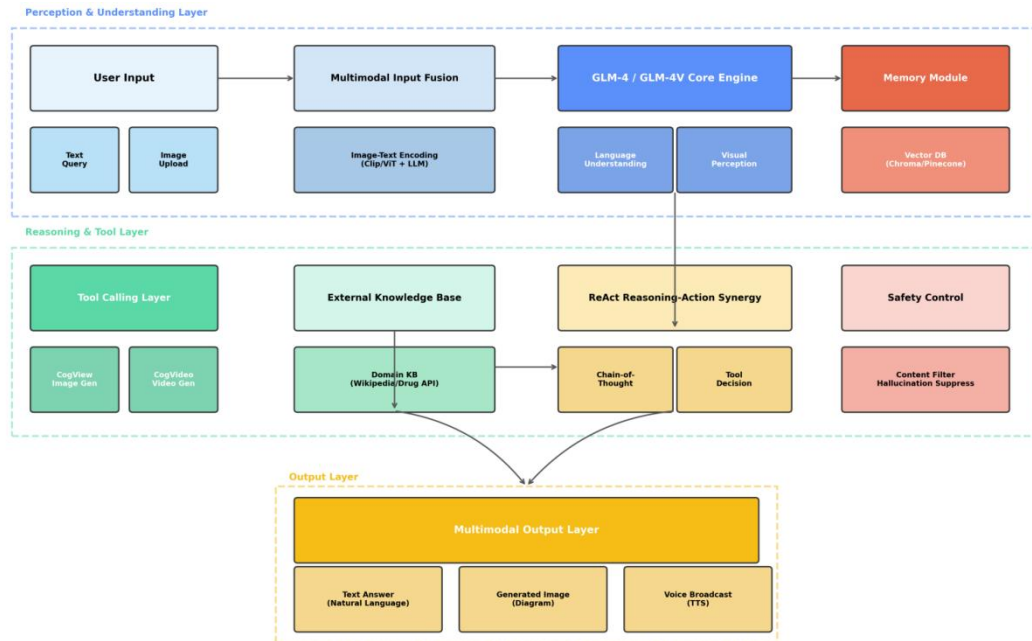


Figure 2 Multimodal Agent System Architecture Based on GLM-4

3.1 ReAct

Alternates Thought and Action with tool calling, see Figure 3 [7]. Advantages: interpretability, error recovery, generalization. 3.2 Memory: working (context window), short-term (summarization), long-term (vector DB). MAG via RAG improves long-document understanding. 3.3 Tool Calling: Function Calling transforms natural language to structured API calls. GLM-4 supports native Function Calling with JSON Schema. 3.4 Multimodal: GLM-4V for understanding, CogView/CogVideo for generation. Config: glm-4v, temp=0.3, max_tokens=512. Results: 84% accuracy, 2.3s response, 0.12 RMB/call [8].

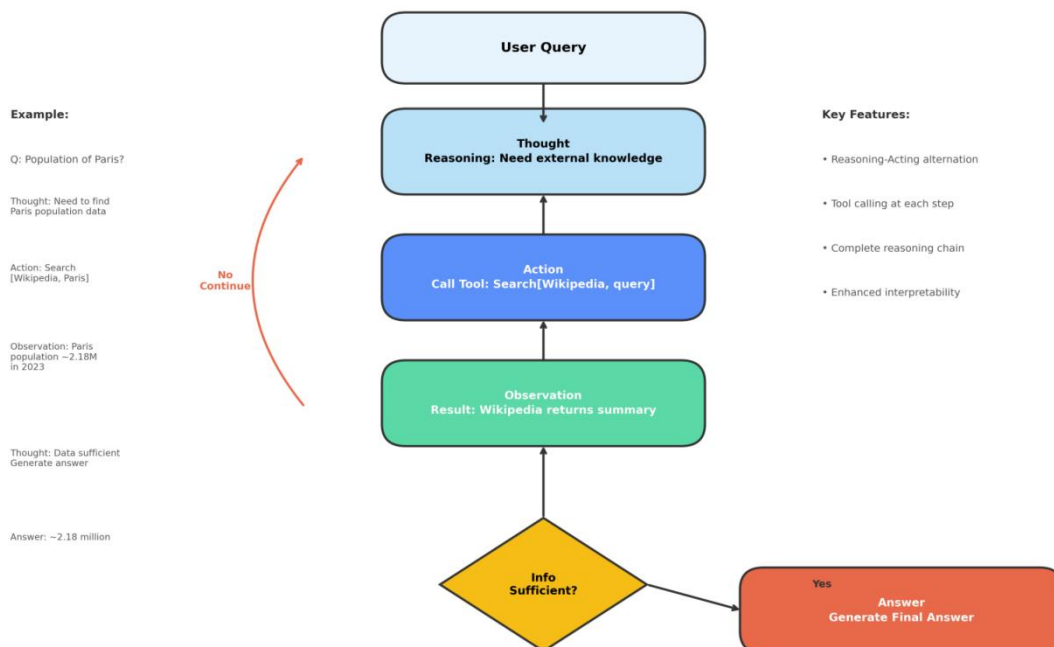


Figure 3 ReAct Reasoning-Acting Synergy Mechanism

4 VERTICAL DOMAIN APPLICATIONS

4.1 Education

Multimodal tutoring for junior high students, see Figure 4 and 5 [9]. Architecture: Student Client → OCR/Whisper → GLM-4V → CogView → Multimodal Output. Results vs traditional: accuracy 84% vs 72%, hallucination 6% vs 12%, satisfaction 4.5/5 vs 3.2/5.

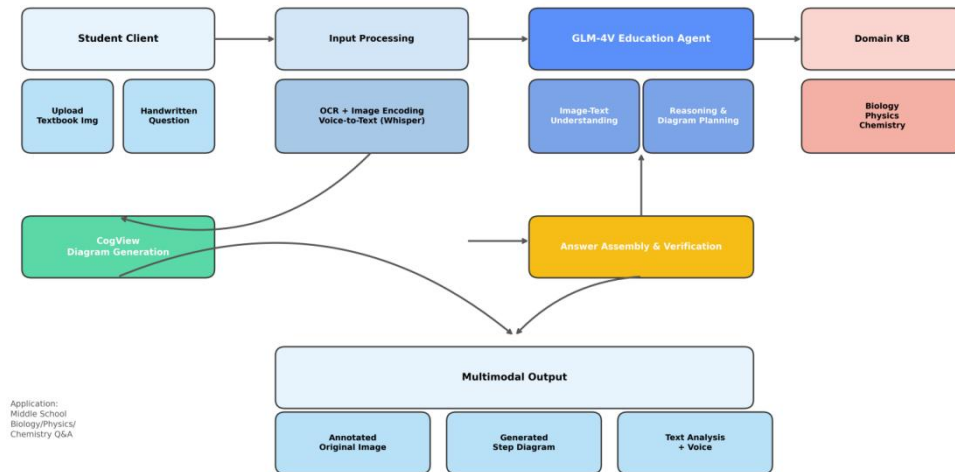


Figure 4 Multimodal Education Assistant System Architecture

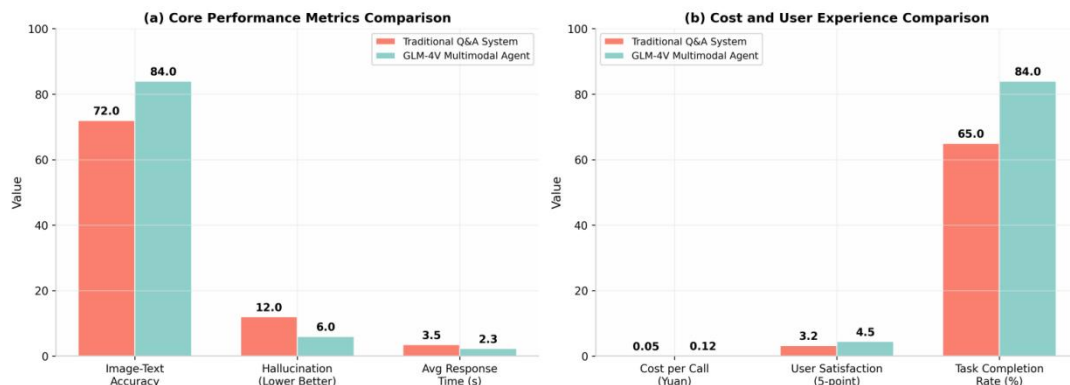


Figure 5 Multimodal Education Assistant Experimental Results

4.2 Healthcare

MedGuard for elderly medication management [10]. Functions: reminders, drug queries via RxNorm API, voice interaction, emergency response, see Figure 6 and 7. Results: on-time rate 60%→95%, satisfaction 4.6/5.0, false alarm 3%, emergency alert success.

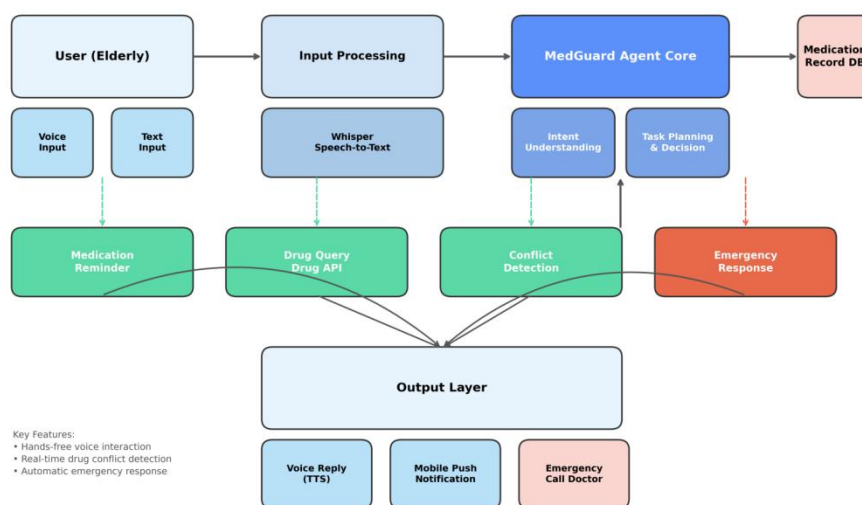


Figure 6 MedGuard System Architecture

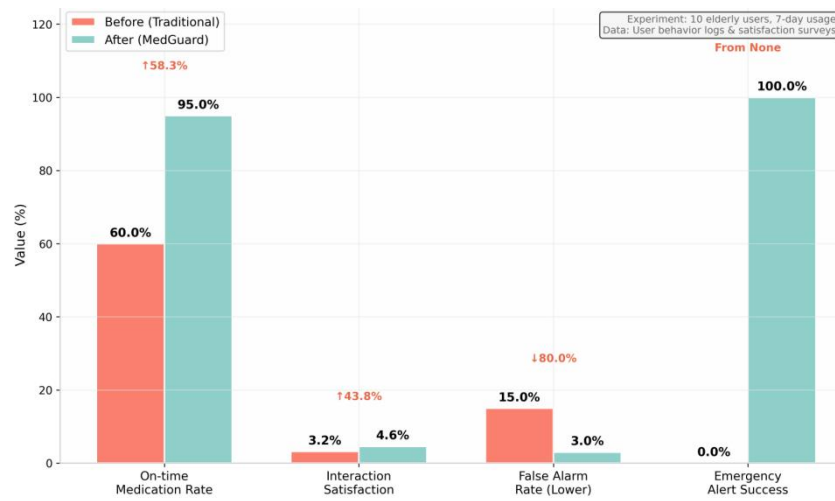


Figure 7 MedGuard Experimental Results

4.3 Multi-Agent Collaboration

MetaGPT [11], AutoGen, CAMEL [12]. Roles: Orchestrator, Planner, Executor, Retriever, Verifier. Shared memory pool for information exchange, see Figure 8.

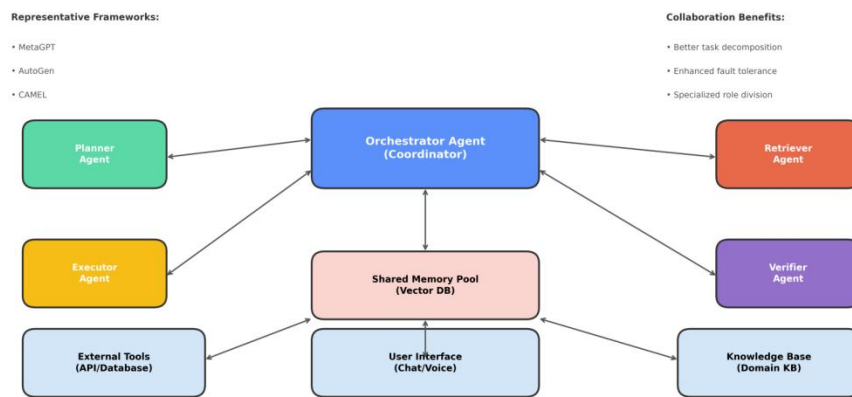


Figure 8 Multi-Agent Collaboration Framework

5 CHALLENGES AND FUTURE DIRECTIONS

Five directions: (1) World Models for causal reasoning; (2) RLHF/RLAIF for alignment [13,14]; (3) Verifiable execution for high-risk actions; (4) Lightweight edge deployment with SLMs; (5) Ethics and legal frameworks with agent audit logs, see Figure 9.



Figure 9 Core Challenge Assessment Radar Chart

6 CONCLUSION

Large model agents represent AI's critical leap from understanding to participating in the world. We reviewed evolution, analyzed ReAct/memory/tools/multimodal technologies, and validated education and healthcare applications. Challenges remain severe: hallucination, alignment, security, cost, interpretability. Future research should focus on world models, RL fine-tuning, verifiable execution, and lightweight deployment [15]. As Sam Altman said: "Agents will be the next platform for AI."

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

AUTHOR CONTRIBUTIONS

Shoutong Huang, Yu Ma, Tong Ding and Rui Ma contributed to the study conception and design, material preparation, data collection and analysis. The first draft of the manuscript was written by Shoutong Huang. All authors commented on previous versions of the manuscript and approved the final version.

REFERENCES

- [1] Du Zhengxiao, Qian Yujie, Liu Xiao, et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). 2022. DOI: 10.18653/v1/2022.acl-long.26.
- [2] Wang L, Ma C, Feng X, et al. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*. 2024, 18(6): 1-26. DOI: 10.1007/s11704-024-40231-1.
- [3] Brooks R A. A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*. 1986, 2(1): 14-23.
- [4] Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. 2nd ed. MIT Press, 2018.
- [5] Anthropic. Introducing the Model Context Protocol. 2024. Available at: <https://www.anthropic.com/news/model-context-protocol>.
- [6] Hong W, Ding M, Zheng W, et al. CogVideo: Large-scale Pretraining for Text-to-Video Generation. In: Proceedings of the International Conference on Learning Representations (ICLR 2023). 2023.
- [7] Yao S, Zhao J, Yu D, et al. ReAct: Synergizing Reasoning and Acting in Language Models. In: Proceedings of the International Conference on Learning Representations (ICLR 2023). 2023.
- [8] Kapoor S, Stroebel B, Siegel Z S, et al. AI Agents That Matter. *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2407.01502.
- [9] Ming Li. Applications and Challenges of Multimodal Large Models in Educational Scenarios. *China Distance Education*, 2025.
- [10] Team GLM. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2406.12793.
- [11] Hong S, Zhuge M, Che J, et al. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *arXiv preprint*, 2023. DOI: 10.48550/arXiv.2308.00352.
- [12] Park J S, O'Brien C, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (ACM CHI 2023). 2023: 1-22. DOI: 10.1145/3586183.3606763.
- [13] Cemri M, Pan M Z, Yang S, et al. Why Do Multi-Agent LLM Systems Fail? *arXiv preprint*, 2025. Available at: <https://arxiv.org/abs/2503.13657>.
- [14] Amodei D. AI Safety in the Age of Agentic Systems. Anthropic, 2024. Available at: <https://www.anthropic.com/news/ai-safety-in-the-age-of-agentic-systems>
- [15] Xi Z, Chen W, Guo X, et al. The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint*, 2023. DOI: 10.48550/arXiv.2309.07864.